Toward Exploring the Role of Disfluencies from an Acoustic Point of View: A New Aspect of (Dis)continuous Speech Prosody Modelling

György Szaszák[1(✉)] and András Beke[2]

[1] Department of Telecommunications and Media Informatics, Budapest University of Technology and Economics, Budapest, Hungary
szaszak@tmit.bme.hu
[2] Research Institute for Linguistics, Hungarian Academy of Sciences, Budapest, Hungary
beke.andras@mta.nytud.hu

Abstract. Several studies use idealized, fluent utterances to comprehend spoken language. Disfluencies are often regarded to be just a noise in the speech flow. Other works argue that fragmented structures (disfluencies, silent and filled pauses) are important and can help better understanding. By extending the original concept of speech disfluency, the current paper involves the acoustic level and places the discontinuity of F0 in parallel with speech disfluencies. An exhaustive analysis of the advantages and disadvantages of using a continuous F0 estimate in prosodic event detection tasks is performed for formal and informal speaking styles. Results suggest that unlike in read (formal) speech, using a continuous, overall interpolated F0 curve is counterproductive in spontaneous (informal) speech. Comparing the behaviour of speech disfluencies and the effect of discontinuity of the F0 contour, results raise more general modelling philosophy considerations, as they suggest that disfluencies in informal speech may be by themselves informative entities, reflected also in the acoustic level organization of speech, which suggests that disfluencies in general are an important perceptual cue in human speech understanding.

Keywords: Disfluency · Interpolation · Prosody · Spontaneous speech

1 Introduction

Spontaneous speech tends to have incomplete syntactic, and even ungrammatical structure and is characterized by disfluencies, repairs and other non-linguistic vocalizations, etc. In general, spontaneous speech is hard to treat with conventional methods developed primarily for the formal or read speaking style, i.e. simple rule-based pattern learning and also data-driven approaches raise several difficulties. One of the most critical challenges is simply determining a broad segmentation of spontaneous speech, such as segmenting speaker turns into utterances.

P. Král and V. Matoušek (Eds.): TSD 2015, LNAI 9302, pp. 369–377, 2015.
DOI: 10.1007/978-3-319-24033-6_42

Sometimes agreement is also missing on the definition of prosodic categories [1]. In current work we rely on the prosodic hierarchy described in [2] and use the phonological phrase level for modelling. We define phonological phrases as prosodic units characterized by an own stress and followed with a less or more complete intonation contour.

Several phrasing or boundary detection approaches have been developed and analysed for read and slightly spontaneous speech (such as semi-formal speech used in information retrieval systems) [3], [4]. The automatic phrasing implemented for read speech in [5] was able to yield a reliable (accuracies ranging between 70-80%) phrasing down to the phonological phrase level, and also to separate intonational phrase level from the underlying phonological phrase level. This approach required clustering of a number of phonological phrase prototypes, which were then modelled by Hidden Markov Models/Gaussian Mixture Models (HMM/GMM) based on acoustic-prosodic features. Clustering of such characteristic prototypes in spontaneous speech was less successful: an effort to try to identify and cluster characteristic prosodic entities or phrase types in Hungarian spontaneous speech by using an unsupervised approach has lead only to partial success [6].

Beside recognizing spontaneous speech as a standalone phenomenon, requiring quite a different approach as compared to read speech, it has been a common practice to try to trace back the processing of spontaneous speech to that of read speech. In other words, use and adapt algorithms or tools developed for read speech in tasks involving processing of spontaneous speech. A characteristic phenomenon, which is often in the focus of this "de-spontaneisation" is speech disfluency, heavily present in spontaneous speech. However, detecting and eliminating or "repairing" disfluency might be counterproductive on some levels of speech processing. Several studies [7] [8] argue that if disfluencies are available in the speech transcription, those can play an important role in disambiguation between sentence-like units. The same can be the case on the acoustic level: preserving disfluency may be sometimes useful. In the current paper, the authors would like to focus on this aspect and compare read and spontaneous speech processing in a prosodic event detection related task.

A frequent disfluency type is filled pause in spontaneous speech. Cook and Lallijee suggested [9] that filled pauses may have something to do with the listener's perception of disfluent speech. They showed that speech may be more comprehensible when it contains filler material during hesitations by preserving continuity and that filled pauses may serve as a signal to draw the listeners attention to the next utterance in order to help the listener not to be able to perceive the onset of the following utterance. Similarly, Swerts and Ostendorf found [10] that in human-machine interactions, turns introducing a new topic tended to have more disfluencies than other turns, showing that a speech recognizer may exploit these disfluencies to detect discourse structure. Swerts et al. analysed the role of filled pauses in discourse structure [11]. They found that phrases following major discourse boundaries contain filled pauses more often.

The silent or filled pauses are generally located at syntactic or prosodic boundaries [12]. Hirst and Cristo demostrated that silent and filled pauses constitute the acoustic markers that enclose the prosodic units [13]. Silent pauses always involve the disruption of the F0 contour. Filled pauses on the other hand are often schwa-like hesitations which are hence voiced. Hereafter, a pitch reset may call the listeners attention and signal the prosodic boundary.

This paper focuses on exploring the advantages and disadvantages of continuity vs discontinuity in speech processing. Recently, several pitch trackers providing stable and overall continuous F0 estimate have been released [14]. Using a continuous, overall interpolated F0 estimation is compared to a case where F0 is kept fragmented or interpolated only partially. The effect of continuous vs discontinuous F0 is evaluated in a prosodic event detection task, separately for read and spontaneous speaking styles. The authors believe that beside yielding some basic, but practically important results these experiments may contribute to a better understanding of spontaneous speech.

The paper is organized as follows: first material (read and spontaneous speech corpora) are presented shortly, followed by the description of the phonological phrase segmentation algorithm used for evaluation in the experiments. Thereafter, experiments are run with original (undefined F0 for unvoiced frames), partially and overall interpolated F0 processing are presented and evaluated for read and spontaneous speech. Finally, conclusions are drawn.

2 Material and Methods

This section describes speech databases and basic processing tools used for the experiments later.

2.1 Speech Databases

BABEL is a Hungarian read speech database, designed for research [15]. A subset of BABEL is used in current experiments, which is labelled for intonational (IP) and phonological phrases (PP). The used subset contains 300 sentences uttered by 22 speakers, containing 2067 PPs, labelled according to 7 types as described in [5].

BEA is a spoken language database [16]. It is the first Hungarian database of its kind in the sense that it involves many speakers, very large amount of spontaneous, informal speech material. The recording conditions of the database were kept permanent and of studio quality. 8 spontaneous narratives were selected (4 male and 4 female) from the database. The sub-corpus was manually annotated by two different phoneticians. The annotation contained three levels: intonational phrases (IP), phonological phrases (PP) and also involved a word level transcription [2]. The IP can be thought of being a part of speech forming a unity in terms of stress and intonation contour, and is found often between two pauses. The database, in most cases the IP boundaries are bound to pauses. A number of filled pauses were perceived as separate IPs by the annotators.

Therefore, filled pauses (FP) were annotated separately in the transcription. As already explained, the IPs can be further divided into PPs based on intonation and stress pattern. A PP is a unity characterized by its own stress and intonation contour, but this latter can be unterminated (continued in next PP). The corpus contained 398 IPs and 751 PPs in total.

2.2 Automatic Segmentation for Phonological Phrases

This section describes the automatic PP segmentation algorithm used in the experiments. Being a prosodic event detection task, its behaviour is analysed with continuous and fragmented F0 patterns. Experiments are run in Hungarian language for read and spontaneous utterances separately.

PPs constitute a prosodic unit, characterized by an own stress and some preceding/following intonation contour. As this contour is specific, PPs can be classified and hence modelled separately, in a data-driven machine learning approach. The distinction between PPs consists of two components: the strength of stress the PP carries and the PPs? intonation contour. In this way 7 different types are distinguished. Modelling is done with HMM/GMM models, and PP segmentation is carried out as a Viterbi alignment of PPs for the utterances requiring segmentation. During this Viterbi alignment, all PPs are allowed to occur with equal probability. A parameter influencing insertion likelihood for PPs can be tuned to force or prevent a more dense segmentation for PPs. The more dense alignment we require, the higher the probability is the insertion of false PP boundaries, resulting often from a confusion between microprosodic variations and accents/prominences resulting from stress. The overall approach is documented in details in [5].

As acoustic-prosodic features, fundamental frequency (F0) and wide-band energy (E) are used [17]. Syllable duration is not used for Hungarian as it was not found to be a distinctive cue in this task [5]. Post-processing alternatives for F0 are described in the respective section later. For energy computation a standard integrating approach is applied with a window span of 150 ms. Frame rate is 10 ms. First and second order deltas are appended to both F0 and E streams.

Evaluation of PP segmentation is done with a 10-fold cross-validation. The PP alignment is generated with models trained on utterances different from the one under segmentation. The generated PP alignment is then compared to the reference obtained by hand-labelling. Detection is regarded to be correct if the boundary is detected within the TOL=100 ms vicinity of the reference. Once all utterances have the automatic PP segmentation ready, the following performance indicators are evaluated:

- recall (RCL) of PP boundaries,
- precision (PRC) of PP boundaries and
- the average time deviation (ATD) between the detected and the reference PP boundary.

3 Overall vs Partial Interpolation of F0

In this section, several scenarios are evaluated using globally or partially continuous $F0$ contours per utterance in the PP alignment task.

All extracted F0 contours are subject to error correction resulting from octal halving/doubling, done by signal processing tool described in [5].

Further post-processing of $F0$ varies according to the scenario, whereas energy is always kept unchanged (by nature continuous, energy is extracted by a 25 ms window each 10 ms, however smoothed with a mean filter of a span of 150 ms afterwards).

3.1 Post Processing Alternatives for F0

The basic interest is to see whether using a continuous contour (all vacancies interpolated) outperforms a non interpolated or partially interpolated contour in the PP detection task. Tested scenarios cover 3 cases as follows:

- Use the F0 contour as produced by a conventional pitch tracker (Snack V2.2.10 in our case [17], doubling/halving errors corrected automatically);
- Use a continuous F0 contour, interpolated at all unvoiced parts;
- Use a partially F0 interpolated contour, where interpolation is omitted if the length of the unvoiced interval exceeds a limit (250 ms in the experiments) or if $F0$ starts significantly higher (suspected pitch reset) than it was before the unvoiced segment (criterion applied in the experiments: $F0_{former} * 1.1 < F0_{current}$).

The motivation to constrain the disruption of the F0 contour in the partial interpolation scenario comes from the following considerations:

- The silence limit is set because a silent period longer than 250 ms can hardly be considered as fluent speech. In such cases the speaker may not employ a pitch reset as the silence in itself can be a clear acoustic marker of PP (and IP) boundary. If this happens interpolation may mask the PP boundary, although energy features are still likely to signal it.
- Medium strength pitch resets may often be smoothed by the F0 interpolation, which makes further detection more difficult. However, a factor of 1.1 is preferred in order to avoid that microprosodic disturbances give false PP (IP) boundary detection, which may happen in the vicinity of long plosives for example.

3.2 Results

Results are shown in Table 1. Regarding precision (PRC) and recall (RCL), they highly depend on a parameter influencing PP insertion likelihoods during the PP segmentation done with Viterbi alignment (see the PRC-RCL curve in Fig. 1). Therefore, segmentation results are shown for operating points where precision and recall are equal (PRC=RCL). The settings of the tolerance interval TOL

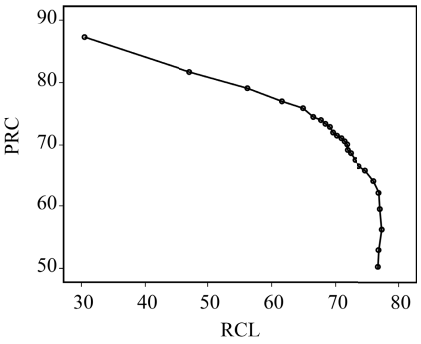

Fig. 1. Precision [%] and recall [%] as influenced by the PP insertion likelihood in the automatic PP segmentation. Read speech, total F0 interpolation, $TOL = 100ms$.

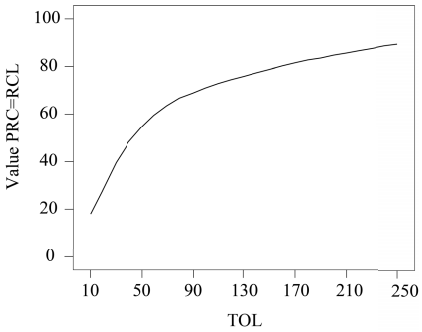

Fig. 2. Precision and recall in operating points defined by $PRC = RCL$ [%] depending on TOL [ms]. Read speech, total F0 interpolation.

Table 1. Precision (PRC) and recall (RCL) in operating points defined by $PRC = RCL$ and ATD for the 3 evaluated scenarios and for read and spontaneous speech ($TOL = 200\ ms$).

Style	F0 interpolation	$PRC = RCL$ [%]	ATD [ms]
Read speech	None	62.9	93.0
	Partial	68.2	80.1
	Total	81.2	55.9
Spontaneous	None	57.7	52.9
	Partial	69.7	44.1
	Total	66.3	44.8

also influence results (see Fig. 2). Unless written else explicitly for the experiments, $TOL = 200ms$ will be used, corresponding roughly to the length of a syllable on average. We believe this deviation is admissible in a supra-segmental detection task, since average PP length is of 659.0 ms in the read speech BABEL corpus and of 782.9 ms in the BEA spontaneous corpus.

As it can be seen from results in Table 1 read and spontaneous speech styles show different behaviour. Whereas in read speech, the more continuous the contour is the better the PP detection results are, this is not the case for spontaneous speech, where the best performing approach constitutes a compromise between do not interpolate at all and interpolate everything: the partially interpolated contour yields the best results for spontaneous speech, where interpolation is omitted if the length of the unvoiced interval exceeds 250 ms or if $F0$ starts by $F0_{former} * 1.1 < F0_{current}$ higher (suspected pitch reset) than it was before the unvoiced segment.

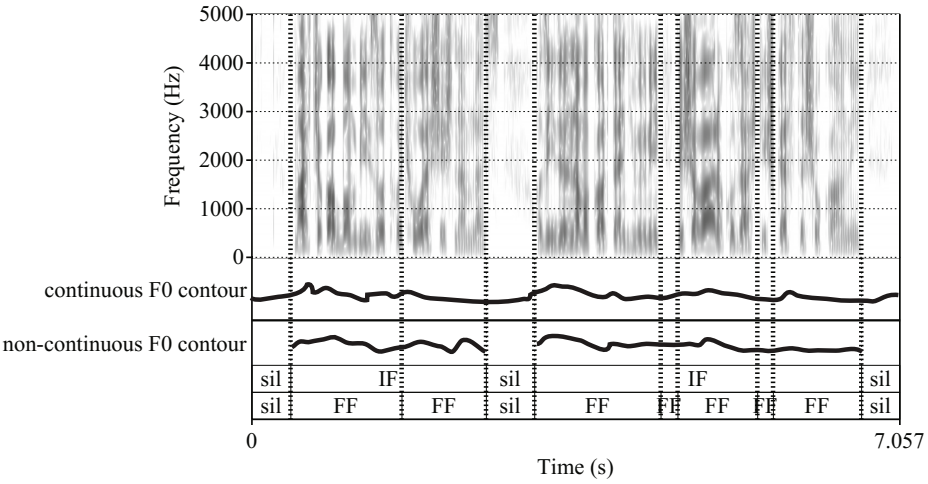

Fig. 3. An example with continuous and partially interpolated F0 contours with IP and PP labelling.

These results suggest that the discontinuity of F0 plays an important cue is spontaneous speech in human perception as well once automatic approaches can exploit it. On the other hand, these results may also make doubtful any attempt to try to "de-spontanize" spontaneous speech in order to transform it into "read" style, and treat it with tools developed for read speech.

4 Conclusions

In the present paper we investigated the effects of using continuous (overall interpolated) vs a fragmented, eventually partially interpolated F0 estimate. Although a noticeable tendency of nowadays is to favour pitch trackers yielding a totally continuous F0 estimate [14], results have shown that this is useful only in read, formal speaking styles, where the overall continuous F0 contour outperformed the partially interpolated one by 19.1% relative in the precision of a phonological phrase segmentation task. According to the results, a partially interpolated F0, where interpolation leaves intact places with longer unvoiced periods or pitch reset suspect F0 increase from one voiced segment to the other, yields by 5.1% relative better results over total interpolation in spontaneous speech in the same PP segmentation task. Beside speech technology applications where spontaneous speech seems to be better treated with only a piecewise interpolation of F0, results also suggest some other considerations regarding human speech perception, however, these latter remain to be confirmed with targeted experiments.

Acknowledgments. The authors would like to thank the support of the Hungarian Scientific Research Fund (OTKA) under contract ID *PD 112598*, titled *"Automatic Phonological Phrase and Prosodic Event Detection for the Extraction of Syntactic and Semantic/Pragmatic Information from Speech"* and the support of the Swiss National Science Foundation via the joint research project *"SP2: SCOPES Project on Speech Prosody"* (No. CRSII2-147611 / 1).

References

1. Silverman, K.M., Beckman, J., Pitrelli, M., Ostendorf, C., Wightman, P., Price, J.P., Hirschberg, J.: Tobi: a standard for labelling english prosody. In: Proceedings of the 2nd International Conference on Spoken Language Processing (ICSLP-92), pp. 867–870 (1992)
2. Selkirk, E.: The syntax-phonology interface. In: International Encyclopaedia of the Social and Behavioural Sciences, pp. 15407–15412. Pergamon, Oxford (2001)
3. Veilleux, N., Ostendorf, M.: Prosody/parse scoring and its application in atis. In: Proceedings of the Workshop on Human Language Technology, pp. 335–340 (1993)
4. Gallwitz, F., Niemann, H., Nöth, E., Warnke, W.: Integrated recognition of words and prosodic phrase boundaries. Speech Communication **36**(1–2), 81–95 (2002)
5. Szaszák, G., Beke, A.: Exploiting prosody for automatic syntactic phrase boundary detection in speech. Journal of Language Modeling **0**(1), 143–172 (2012)
6. Beke, A., Szaszák, G.: Unsupervised clustering of prosodic patterns in spontaneous speech. In: Sojka, P., Horák, A., Kopeček, I., Pala, K. (eds.) TSD 2012. LNCS, vol. 7499, pp. 648–655. Springer, Heidelberg (2012)
7. Medeiros, H., Batista, F., Moniz, H., Trancoso, I., Meinedo, H.: Experiments on automatic detection of filled pauses using prosodic features. Actas de Inforum **2013**, 335–345 (2013)
8. Swerts, M.: Filled pauses as markers of discourse structure. Journal of Pragmatics **30**, 485–946 (1998)
9. Cook, H., Lallijee, M.: The interpretation of pauses by the listener. Brit. J. Soc. Clin. Psy. **9**, 375–376 (1970)
10. Swerts, M., Ostendorf, M.: Prosodic and lexical indications of discourse structure in human-machine interactions. Speech Communication **22**(1), 25–41 (1997)
11. Swerts, A., Wichmann, A., Beun, R.J.: Filled pauses as markers of discourse structure. In: Proceedings ICSLP96, Fourth International Conference on Spoken Language Processing, pp. 1033–1036 (1996)
12. Zellner, B.: Pauses and the temporal structure of speech. In: Fundamentals of Speech Synthesis and Speech Recognition, pp. 41–62. John Wiley, Chichester (1994)
13. Hirst, D., Cristo, A.D.: Intonation Systems: A Survey of Twenty Languages. Cambridge University Press, New York (1989)
14. Ghahremani, P., BabaAli, B., Povey, D., Riedhammer, K., Trmal, J., Khudanpur, S.: A pitch extraction algorithm tuned for automatic speech recognition. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 2494–2498 (2014)
15. Roach, P.S., Amfield, S., Bany, W., Baltova, J., Boldea, M., Fourcin, A., Goner, W., Gubrynowicz, R., Hallum, E., Lamep, L., Marasek, K., Marchal, A., Meiste, E., Vicsi, K.: Babel: an eastern european multi-language database. In: International Conf. on Speech and Language, pp. 1033–1036 (1996)

16. Neuberger, T., Gyarmathy, D., Gráczi, T.E., Horváth, V., Gósy, M., Beke, A.: Development of a large spontaneous speech database of agglutinative Hungarian language. In: Sojka, P., Horák, A., Kopeček, I., Pala, K. (eds.) TSD 2014. LNCS, vol. 8655, pp. 424–431. Springer, Heidelberg (2014)
17. Sjölander, K., Beskow, A.: Wavesurfer - an open source speech tool. In: Proceedings of the 6th International Conference of Spoken Language Processing, vol. 4, pp. 464–467 (2000)